(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 09034651 A
(43) Date of publication of application: 07.02.1997

(51) Int. Cl G06F 3/06

(21) Application number: 07183841
(22) Date of filing: 20.07.1995

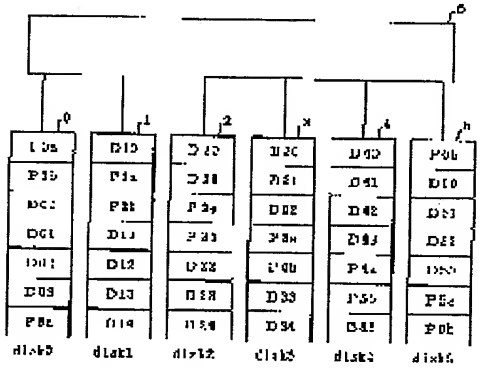
(71) Applicant: HITACHI LTD
HITACHI CHUBU SOFTWARE LTD
(72) Inventor: HAYASHI ITSUKI
FUJIOKA YOSHINORI
FUKAYA YASUKATSU
FUKUMOTO SATOSHI

(54) DISK ARRAY DEVICE

(57) Abstract:

PROBLEM TO BE SOLVED: To reduce disk contention accompanying. parity update and to improve the performance of a disk array by forming the same parity group by first and second parities and updating only. the first or second parity at the time of parity update.

SOLUTION: This device consists of disk devices 0 to 5 (disk 0 to disk 5) of units and an array control mechanism 6. Each of disk devices 0 to 5 is divided into blocks as sets of plural sectors. In this constitution of distributed sparing, spare areas are used as second parity areas, and second parities (POb, P1b...) forming the same parity groups as first parities (POa, P1a...) are stored there. Second parities are provided in the disk array in this manner, and first and second parities are arranged in the same parity group, and only one of two parities is updated.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平9－34651

(43)公開日　平成９年(1997)２月７日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　3/06	５４０		Ｇ０６Ｆ　3/06	５４０	

審査請求　未請求　請求項の数７　ＯＬ　（全 ６ 頁）

(21)出願番号　特願平7－183841

(22)出願日　平成７年(1995)７月20日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(71)出願人　000233457
日立中部ソフトウェア株式会社
愛知県名古屋市中区栄３丁目10番22号

(72)発明者　林　逸樹
愛知県名古屋市中区栄三丁目10番22号日立中部ソフトウエア株式会社内

(72)発明者　藤岡　良記
愛知県尾張旭市晴丘町池上１番地株式会社日立製作所オフィスシステム事業部内

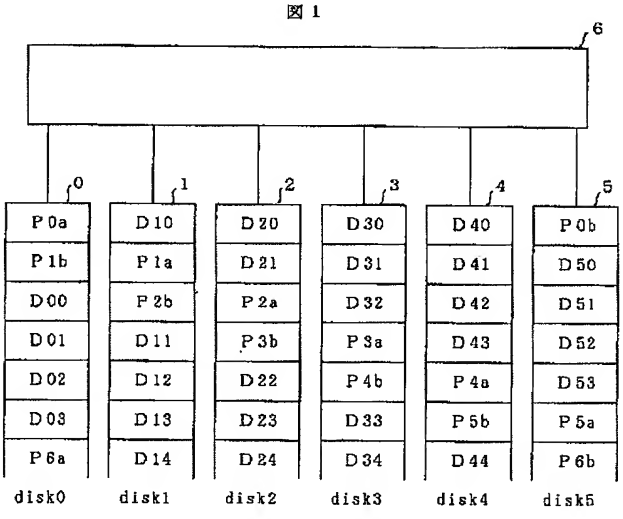
(74)代理人　弁理士　小川　勝男

最終頁に続く

(54)【発明の名称】　ディスクアレイ装置

(57)【要約】
【目的】スペアディスクを常備するディスクアレイの性能の改善を図る。
【構成】ディスクアレイ内に第１、第２の複数のパリティを設け、第１のパリティおよび第２のパリティを同一のパリティグループに配置し、パリティの更新を２つのパリティのうちの１つに対してのみ行う。障害回復時には、どちらかをスペア領域にする。
【効果】パリティ競合によるオーバヘッドを軽減でき、ディスクアレイの性能を改善できる。

図１



1

【特許請求の範囲】

【請求項１】アレイ構成の複数台のディスクドライブとアレイ制御機構からなるディスクアレイ装置であって、該各ディスクドライブは複数のセクタの集まりであるブロックに分割され、データおよびパリティがブロック単位にアレイ内に分散され、各ブロックが独立にアクセスが可能なディスクアレイ装置において、複数のデータおよびパリティとのパリティ計算単位であるパリティグループ内に２以上の複数パリティを含ませたことを特徴とするディスクアレイ装置。

【請求項２】請求項１のディスクアレイ装置において、前記アレイ制御機構に対するライト要求におけるパリティ更新を前記複数のパリティのうちの１つのみを更新することにより実行することを特徴とするディスクアレイ装置。

【請求項３】請求項１または２記載のディスクアレイ装置において、前記複数台のディスクドライブのうちの１台が故障した場合は、前記複数のパリティのうちの１つをスペア領域として、消失データおよび新パリティの生成を行い、リカバリを実行することを特徴とするディスクアレイ装置。

【請求項４】アレイ構成の複数台のディスクドライブとアレイ制御機構からなるディスクアレイ装置であって、該各ディスクドライブは複数のセクタの集まりであるブロックに分割され、各ブロックが独立にアクセスが可能なディスクアレイ装置において、前記複数台のディスクドライブのうちの２以上の複数台をパリティディスクとし、該複数台のパリティディスクのパリティを同一のパリティグループ内に配置したことを特徴とするディスクアレイ装置。

【請求項５】請求項４記載のディスクアレイ装置において、前記アレイ制御機構に対するライト要求におけるパリティ更新を前記複数台のパリティディスクのうちの１つのみを更新することにより実行することを特徴とするディスクアレイ装置。

【請求項６】請求項４又は５記載のディスクアレイ装置において、前記複数台のディスクドライブのうちの１台が故障した場合は、前記複数台のパリティディスクのうちの１つをスペアディスクとして、消失データおよび新パリティの生成を行い、リカバリを実行することを特徴とするディスクアレイ装置。

【請求項７】請求項６記載のディスクアレイ装置において、前記ディスク１台のディスク故障に対するリカバリ動作の後の、新しいパリティディスクを追加して故障前のアレイ構成に戻すコピーバック動作を、前記追加したパリティディスクの内容をオール'０'にイニシャライズすることにより実行することを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【０００１】

2

【産業上の利用分野】本発明はディスクアレイ装置に関し、特にライトコマンドに対するパリティ更新のオーバヘッドを軽減することの可能なディスクアレイ装置に関する。

【０００２】

【従来の技術】ディスクアレイは、アレイ構成された複数のディスクを並列アクセスすることによりアクセススピード等の性能の向上を図り、データにパリティを付加して、１台のディスク故障に対して消失データの自動回復が可能な、高性能、高信頼性ディスクシステムである。

【０００３】ディスクアレイの基本構成は、Patterson等によりＲＡＩＤ（Redun-dant Arrays of Inexpensive Disks）として５つのレベルにまとめられ、このうち、レベル４（ＲＡＩＤ４）およびレベル５（ＲＡＩＤ５）は、アクセス単位をブロック単位でインターリーブする方法である。中でもＲＡＩＤ５は、パリティをアレイ内に分散することにより、パリティディスクへの負荷分散を可能とした、最も分散処理システム向けの方法と考えられている。

【０００４】ＲＡＩＤ５およびＲＡＩＤ４の性能上の最大の問題点は、パリティ更新のために、ライトコマンドが常にリードモディファイライトアクセスになることである。　即ち、１回のライトコマンドに対して、４回のディスクアクセス（データおよびパリティのリードとデータおよびパリティのライト）が必要になる。特に、ＲＡＩＤ４では、パリティが１台のディスクに集中しているため、パリティを格納しているディスクへのアクセスが性能上のボトルネックになっている。

【０００５】このオーバーヘッドの軽減方法として、スペアリングの方法が提案されている。オンライントランザクション処理等のシステムでは、ディスクが故障した場合の回復処理（リカバリ）をシステムを停止することなく実行するために、通常、ホットスタンバイのスペアディスクを常備している。スペアリングとは、スペアディスクの利用技術のことである。Proceedings of 19th Annual InternationalSymposium on Computer Architecture において、Menon等がこの分散スペアリングの方法を提案し、その性能を評価している。

【０００６】分散スペアリングは図３に示す如く、各ディスク（disk0～disk5）にパリティ（Ｐ0,Ｐ1,Ｐ2…）だけでなく、スペア領域（Ｓ0,Ｓ1,Ｓ2…）も分散させる方法である。このスペア領域をを利用することにより、データディスクの数を実質的に１台増加させることができ、アレイ性能の改善を図っている。

【０００７】

【発明が解決しようとする課題】上記分散スペアリングでは、データディスクの数は増やすことができるが、前述したライトコマンドに対するパリティネックの問題は基本的には改善されていない。即ち、図３において、デ

ータＤ20とパリティＰ0をアクセスするライトコマンド１を実行している時に、データＤ32とパリティＰ2をアクセスするライトコマンド２を実行することはできない。

【０００８】本発明の目的は、上記問題点を解消し、ライトコマンドに対するパリティの更新に伴うディスク競合を軽減させる方法を提供し、ディスクアレイの性能の改善を図ることにある。更に本方法をＲＡＩＤ４にも適用し、ＲＡＩＤ４におけるパリティディスクのボトルネックを大幅に改善することにある。

【０００９】

【課題を解決するための手段】本発明の上記目的は、分散スペアリングの構成において、スペア領域（Ｓ0，Ｓ1，Ｓ2…）を、図１に示す如く第２のパリティ領域として使用し、第１のパリティ（Ｐ0a，Ｐ1a，Ｐ2a…）と同一のパリティグループを成す第２のパリティ（Ｐ0b，Ｐ1b，Ｐ2b…）を格納することにより実現できる。また、スペアディスクを常備するＲＡＩＤ４構成のディスクアレイにおいて、スペアディスクを図２に示す如く第２のパイリティディスクとして使用し、第１のパリティ（Ｐ0a，Ｐ1a，Ｐ2a…）と同一のパリティグループを成す第２のパリティ（Ｐ0b，Ｐ1b，Ｐ2b…）を格納することにより実現できる。

【００１０】

【作用】第１および第２のパリティで同一のパリティグループを形成するため、パリティの更新は第１もしくは第２のどちらかのパリティのみを更新すればよく、パリティ更新に伴うディスク競合を軽減でき、ディスクアレイの性能改善を図れる。

【００１１】また、ディスクの障害回復（リカバリ）時には、第１または第２のパリティをスペア領域として使用するため、ディスクアレイの障害回復機能を維持できる。

【００１２】

【実施例】以下、本発明の一実施例を図面を用いて詳細に説明する。

【００１３】図１は本発明の一実施例をＲＡＯＤ５に適用した例を、図２は本発明の一実施例をＲＡＩＤ４に適用した例をそれぞれ示す。図１及び図２において、０～５（disk0～disk5）は単体のディスク装置、６はアレイ制御機構である。

【００１４】各ディスク装置０～５はそれぞれ、複数のセクタの集まりであるブロックに分割される。それぞれのブロックには、当該ディスク装置に格納するデータがデータブロック（Ｄ01，Ｄ02，Ｄ03…）と、第１のパリティブロック（Ｐ0a，Ｐ1a，Ｐ2a…）、第２のパリティブロック（Ｐ0b，Ｐ1b，Ｐ2b…）に分割されて前記各ディスク装置のブロックに格納される。

【００１５】図４は本発明におけるパリティの更新アルゴリズムの一例を示した図である。

【００１６】図１において、第１のパリティと第２のパリティはディスク０～５内に分散されていて、ディスク０～５の横１列のブロックにより、１つのパリティグループが形成される。即ち、パリティを偶数パリティと仮定すると、

$$P0a \text{ XOR } D10 \text{ XOR } D20 \text{ XOR } D30 \text{ XOR } D40 \text{ XOR } P0b = 0$$

の関係を持たせて格納させる。ここで、XOR は排他的論理和を表す。

【００１７】ディスク０～５（disk0～disk5）の初期化は各パリティグループ毎に行われる。即ち、まず、初期化されたディスクに新たにデータを格納するときには、ディスク装置の横一列のパリティグループについて、４個のデータブロックについて生成したパリティを第１または第２のパリティブロックの一方に書き込み、残りのパリティブロックにオール'０'を書き込んでおく。

【００１８】例えば、図１の最上位のパリティグループにおいて、Ｄ10，Ｄ20，Ｄ30，Ｄ40について生成したパリティをＰ0aに書き込み、Ｐ0bにはオール'０'を書き込む。それに続くデータについても同様の処理を行なって、データを格納しておく。

【００１９】格納されたデータの一部又は全部を書き換える（更新）するときの制御は、図４に示すアルゴリズムに従う。上位システムからアレイ制御機構６にライトコマンドがリクエストされると（処理１１）、アレイ制御機構６は、更新するデータの存在するディスク装置と、更新するデータのパリティグループの第１および第２のパリティの存在するディスク装置が動作中（ビジー）かどうかをチェックする（処理１２，１３）。

【００２０】もし書き換えるデータが存在するディスク装置へのアクセスがビジーの場合は、データのディスクのアクセスが終了するまでライトコマンドの実行をウェイトする。そして、書き換えるデータが存在するディスク装置がビジーでない場合には、当該データのパリティーが存在するディスク装置がビジーか否かをチェックする。　本発明では、書き換える１つのデータブロックに対して２つのパリティーが存在するが、このチェックでパリティが存在するディスク装置が両方共ビジーの場合は、どちらかのディスク装置のアクセスが終了するまでライトコマンドの実行をウェイトする。もし片方のパリティの存在するディスク装置がビジーでない場合には、ビジーでないディスク装置のパリティを用いてパリティの更新を行う（処理１４）。この際には他方のパリティについてはそのままのパリティーとしておく。

【００２１】パリティの存在するディスク装置が両方共にビジーでない場合は、あらかじめ決められたどちらかのパリティ（例えば第１のパリティー）を用いてデータブロック及びパリティの更新を行う（処理１５）。

【００２２】なお、パリティの存在する両方のディスク

装置が共にビジーでない場合に使用するパリティの決め方は、最後のアクセスが古い方を選択する方法などでもよい。

【００２３】例として、今、データＤ20およびパリティＰ0aへのライトコマンド１を実行中に、データＤ32へのライトコマンド２がアレイ制御機構６にリクエストされたとすると、アレイ制御機構６は、パリティＰ2aがビジー中のため、データＤ32およびパリティＰ2bでライトコマンド２を実行する。即ち、データＤ32およびパリティＰ2bを読みだし、該読みだしたデータおよびパリティとデータＤ32の更新データにより、更新パリティを生成し、データＤ32に更新データを書き込むと同時に、パリティＰ2bに該更新パリティを書き込む。

【００２４】図３で示す分散スペアリングでは、パリティＰ2bがスペア領域であり、パリティＰ2aがデータＤ20と競合するため、ライトコマンド２の実行はライトコマンド１が終了するまで待たされることになる。

【００２５】一方図２において、第１のパリティ（Ｐ0a, Ｐ1a, Ｐ2a…）および第２のパリティ（Ｐ0b, Ｐ1b, Ｐ2b…）はそれぞれ専用のディスク装置、ディスク装置４（disk4）、ディスク装置５（disk5）に格納されている。パリティグループは図１と同様の横１列のブロックで構成される。ＲＡＩＤ４では、ディスク装置５（disk5）がスペアディスクのため、ライトコマンドの並列アクセスが不可能であるが、本構成では、２台のパリティディスク（disk4, disk5）の内の１つのパリティを更新すればよいから、ライトコマンドの並列アクセスが可能になる。

【００２６】例えば、データＤ00およびパリティＰ0aに対するライトコマンドとデータＤ11およびパリティＰ1bに対するライトコマンドの並列アクセスが可能になる。

【００２７】図５は図１におけるアレイ制御機構６の内部構成を詳細に示した図である。図５において、０～５はディスク装置、６はアレイ制御機構、７はバスインタフェース、８はリードバッファ、９はアドレス変換機構、１０～１５はディスクコントローラ、１６はマイクロコントローラ、１７はマルチプレクサ、１８はリードモディファイライト用バッファ、１９はＸＯＲロジック、２０はコマンドバッファ、２１はパリティマッピング用テーブルである。上位インタフェース１０６からのリード／ライトコマンドは、バスインタフェース７を経由して、アドレス変換機構９により、各ディスク装置０～５専用のキューイングバッファであるコマンドバッファ２０に割り当てられる。この割当て方法はマイクロコントローラ１６によりプログラマブルに変更できる。

【００２８】マイクロコントローラ１６はライン１０８により、コマンドバッファ２０のコマンド内容を認識できる。リードコマンドが先頭の場合は、マイクロコントローラ１６はライン１１０により、当該ディスク装置がビジーかどうかを判定し、もしビジーでなければ、当該ディスクコントローラにリード起動をかけ、当該ディスク装置からデータを読み込む。該読みだされたデータは、リードバッファ８に格納され、ライン１０７、バスインタフェース７経由で上位インタフェースバス１０６に送出される。

【００２９】一方、ライトコマンドが先頭の場合は、マイクロコントローラ１６はパリティマッピング用テーブル２１により、該ライトコマンドに対応する２つのパリティのアドレスを調べ、該２つのパリティを持つディスクがビジーかどうかを判定し、ビジーでないパリティを更新する。もし、両方共ビジーでなければ、予め決められたぱれてぃを更新する。

【００３０】パリティの更新は以下の手順で行う。マイクロコントローラ１６はライトするブロックの古いデータと、当該パリティブロックの古いパリティを、当該ディスク装置より読出し、該読みだした古いでた及びパリティはマルチプレクサ１７経由でリードモディファイライト用バッファ１８に格納する。

【００３１】次にマイクロコントローラ１６はマルチプレクサ１７経由でコマンドバッファ２０から読みだしたライトデータと、前記リードモディファイライト用バッファ１８に格納した古いデータおよびパリティを、ＸＯＲロジック（排他的論理和）１９に入れ、新しいパリティを生成する。該生成した新しいパリティは、ライン１０９経由で当該ディスクコントローラ経由で、当該ディスク装置に書き込まれる。この時、同時に、ライトデータも当該ディスク装置に書き込まれる。

【００３２】上記リード及びライトコマンドは当該ディスク装置が仕様可能であれば、並列に処理することができる。

【００３３】次に、ディスク装置に故障が生じた後の回復処理について説明する。スペア領域を用いる図３に示す方式の場合、１台のディスク装置が故障した場合、ディスクアレイ装置の制御装置は消失データの回復処理を実行する。この時、再生成したデータはスペア領域に格納する。本実施例では、第１のパリティ（Ｐ0a, Ｐ1a, Ｐ2a…）または第２のパリティ（Ｐ0b, Ｐ1b, Ｐ2b…）のどちらか一方（所定のルールにより予め決めておく）をスペア領域として使用する。また、残りのパリティは、故障ディスクの替わりの新しいディスクが用意されるまで、通常のＲＡＩＤ５またはＲＡＩＤ４構成で動作させるために再生成をする。消失したパリティの再生成は行わない。

【００３４】例えば、図１において、ディスク１（disk1）が故障したとすると、パリティエリアＰ0bに消失したデータＤ10を生成し、パリティＰ0aにはデータＤ20, Ｄ30, Ｄ40, Ｄ10の新しいパリティを生成して格納する。消失したパリティＰ1aは再生成せず、Ｐ1bにデータＤ21, Ｄ31, Ｄ41, Ｄ51の新しいパリティを生成する。図２のＲＡＩＤ４においても、同様の処理を行う。

【００３５】回復処理の終了後、故障ディスクの替わりの新しいディスクが用意されると、ディスクアレイの構成を故障前の状態に戻すコピーバック動作を行なう。

【００３６】本動作は、図１に示すＲＡＩＤ５構成の分散パリティでは、新しいディスクにデータをコピーすると共に、２つのパリティを再生成することにより実行する。

【００３７】一方、図２のＲＡＩＤ４構成の分散パリティでは、パリティが予め定められた２つの決まったディスクに格納されるので、コピーバック動作は行わない（必要ない）。オール'０'にイニシャライズされた新しいディスクを用意し、第２のパリティディスクとするだけでよい。

【００３８】

【発明の効果】以上、説明した如く、本発明によれば、＊ディスクアレイ内に第２のパリティを設け、第１のパリティおよび第２のパリティを同一のパリティグループに配置し、パリティの更新を２つのパリティのうちの１つに対してのみ行うことにより、パリティ競合によるオーバヘッドを軽減し、ディスアレイの性能を改善することができる。

【図面の簡単な説明】

【図１】本発明の第１の実施例を示した図。

【図２】本発明の第２の実施例を示した図。

【図３】従来技術を示した図。

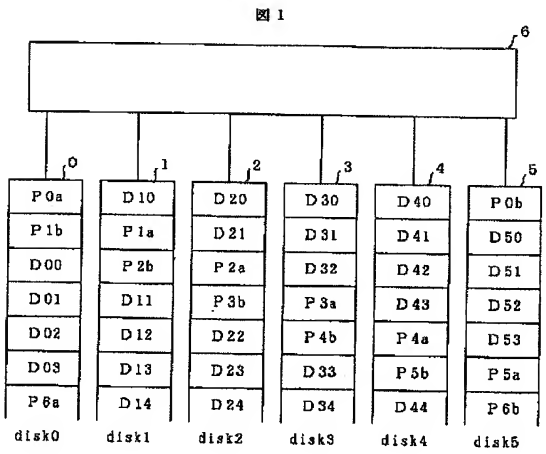
【図４】本発明におけるパリティ更新のアルゴリズムの例を示した図。

【図５】図１の詳細構成を示す図。

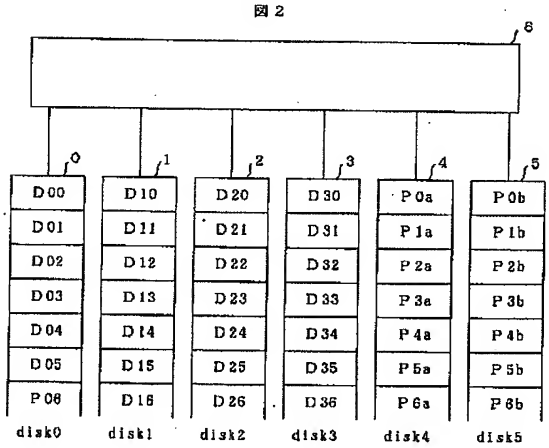
【符号の説明】

０～５…ディスク装置、６…アレイ制御機構。

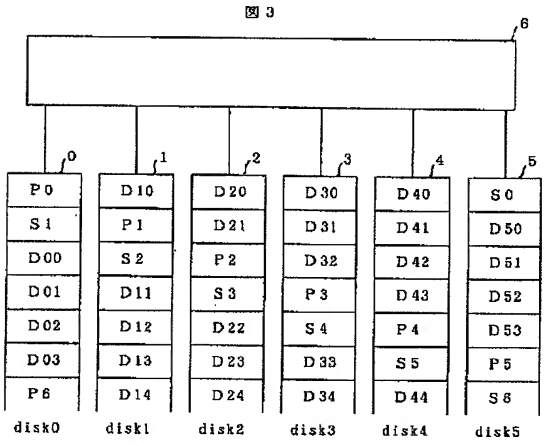
【図１】



【図２】

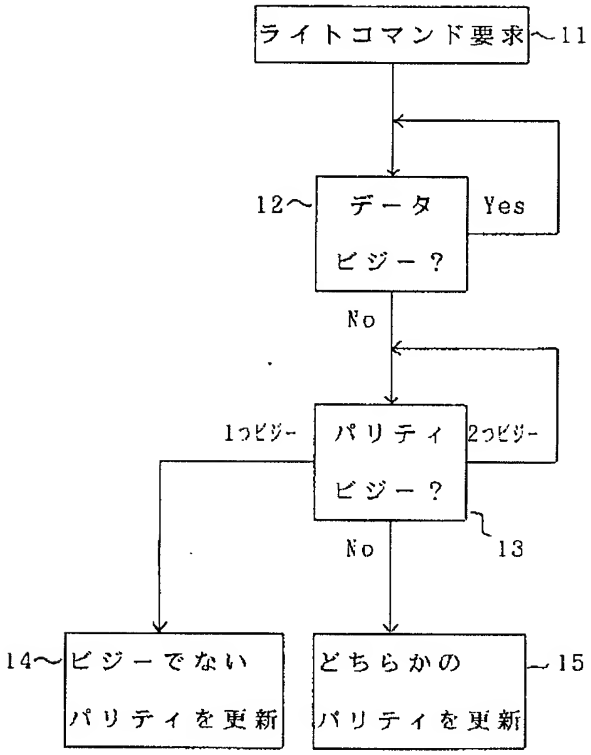


【図３】

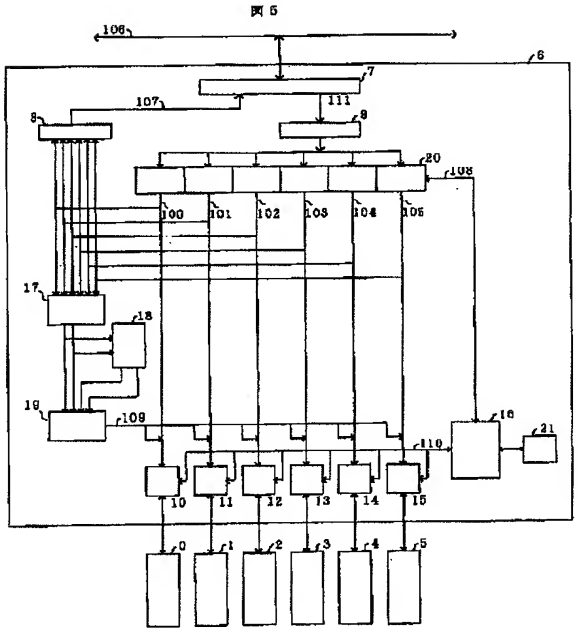


【図４】

図４



【図５】



フロントページの続き

(72)発明者　深谷　寧克
愛知県尾張旭市晴丘町池上１番地株式会社
日立製作所オフィスシステム事業部内

(72)発明者　福本　聡
愛知県豊田市八草町八千草1247